****

**Attrition Prediction Analytics**

**Solution for Deloitte Talent Team**

Document Control Information

Document Information

|  |  |
| --- | --- |
| Document Identification | 17 |
| Document Name | **Attrition Analytics Document** |
| Project Name | **Attrition Analytics** |
| Client | Deloitte |
| Document Author | **Deloitte Team** |
| Document Version | **1.1** |
| Document Status | **Final Document** |
| Date Released |  |

Document Edit History

| Version | Date | Additions/Modifications | Prepared/Revised by |
| --- | --- | --- | --- |
| 1.0 | 17-Sep-20 | SOA | Deloitte Team |
| 1.1 | 18-Sep-20 | SOA | Deloitte Team |
| 1.2 | 23-Sep-20 | SOA | Deloitte Team |
| 1.3 |  |  |  |
| 1.4 |  |  |  |
| 1.5 |  |  |  |
| 1.6 |  |  |  |
| 1.7 |  |  |  |
| 1.8 |  |  |  |
| 1.9 |  |  |  |

Document Review/Approval History

| Date | Name | Organization/Title | Comments |
| --- | --- | --- | --- |
| 18-Sep-20 | Mitra, Indrajit | Deloitte |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Distribution of Final Document

The following people are designated recipients of the final version of this document:

| Name | Organization/Title |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

**Table of Contents**

1. [Project Scope](#ProjectScope)5

2. [Understanding Project Requirement and Planning](#UnderstandProjectRequirementandPlanning)5

3. [Data Collection, Preparation and Derived Variables Creation](#DataCollectPrepDerivedVarCreation)6

3.1 [Identification and Analysis of Data Sources](#IdentificationAndAnalysisofDataSources) 6

3.2 [Data Structure](#DataStructure)7

3.3 [Data Preparation](#DataPreparation) 7

3.4 [Derived Variables Creation](#DerivedVariablesCreation) 8

4. [Model Development](#ModelDevelopment)11

4.1 [Defining Model Selection Metrics](#DefiningModelSelectionMetrics)11

4.2 [Data Strategy for Model Development](#DataStrategyforModelDevelopment)12

4.3 [Model Exploration](#ModelExploration)12

4.4 [Variables Reduction](#VariablesReduction)13

4.5 [Model Refinement](#ModelRefinement)13

4.6 [Model Validation and Testing](#ModelValidationandTesting)14

4.7 [Final Models and Results](#FinalModelsandResults)14

**4.7.1** [**Consulting Model**](#ConsultingModel)**14**

**4.7.2** [**Risk Advisory Model**](#RiskAdvisoryModel)**15**

4.8 [Version Control](#VersionControl)17

4.9 [Lime Explanations](#LimeExplanations)17

4.10 [Model Output](#ModelOutput)18

5. [Model Refresh SOP](Attrition%20Analytics_v1.docx)19

5.1 [Architecture](#Architecture)19

5.2 [Folder Structure](#FolderStructure)21

5.3 [Process Flow](#ProcessFlow)21

5.3.1 [Model Input](#ModelInput)22

5.3.2 [Model Code Run Procedure and Logs](#ModelCodeRunProcedureandLogs)23

5.3.3 [Output Generation](#OutputGeneration)24

6. [Project Assumptions and Controls](#ProjectAssumptionsandControls)24

6.1 [Project Assumptions](#ProjectAssumptions)24

6.2 [Project Controls](#ProjectControls)25

**1. Project Scope:**

* Deloitte talent team is looking for to improve the employee retention strategy and thereby to improve upon the employee attrition
* The project will deliver two outcome upon successful completion
  + Employee attrition risk score
  + Top reasons causing the attrition

Additionally, the talent will be able to generate insights such as to look at consolidated employee information book, comparable insights between two business units, trend attrition by various cuts (region, business units, supervisor, year and month etc.)

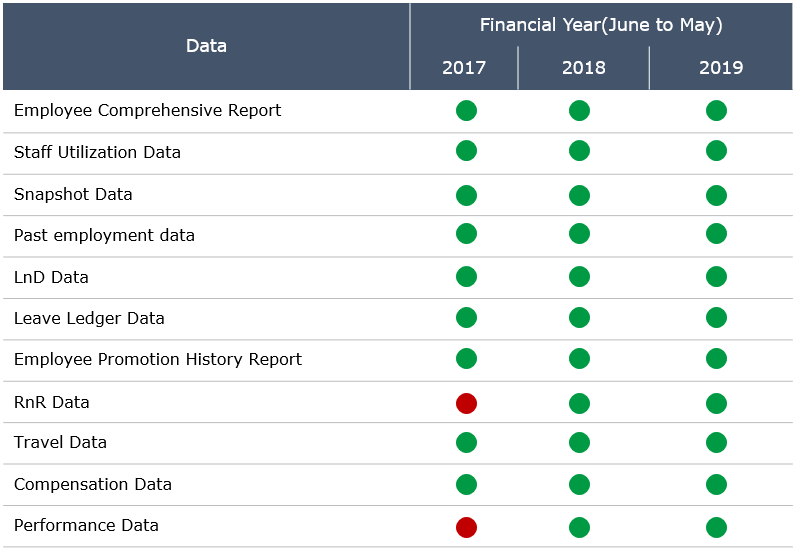
**2.** **Understanding Project Requirement and Planning:**

1. One pilot run is planned that will cover Consulting and RA, as recommended by Talent and Innovation teams
2. The timelines and effort, cover the standard data cleansing, transformation and model development
3. The timeline is valid only for activities as mentioned and data provided by Talent, which has been received and agreed upon by the Development Team
4. Though the pilot covers two BUs, the data preparation will encompass the entire data (for ~17000 observations currently) and future modelling work for the remaining BUs can be performed using this data
5. On infrastructure, an offline mode will be used that will allow to run all activities as mentioned above using Deloitte laptop, Deloitte Cloud (ex. MS Teams/SharePoint), freeware such as R and Python
6. Industrialization including periodic refresh, maintenance and any visualization development is out of scope
7. The Development team will deliver output files as per pre-described format such as CSV, R or Python files along with the model codes
8. The output will be valid for three months from the date of final run to analyze the attrition probabilities
9. Subsequent runs, post the pilot run, may accommodate one or two BUs and can be delivered in a two-three weeks’ time window

**3.** **Data Collection, Preparation and Derived Variables Creation:**

**3.1** **Identification and Analysis of Data Sources:**

Variety of data has been collected to capture the holistic view of the employees. They are listed in the table below:

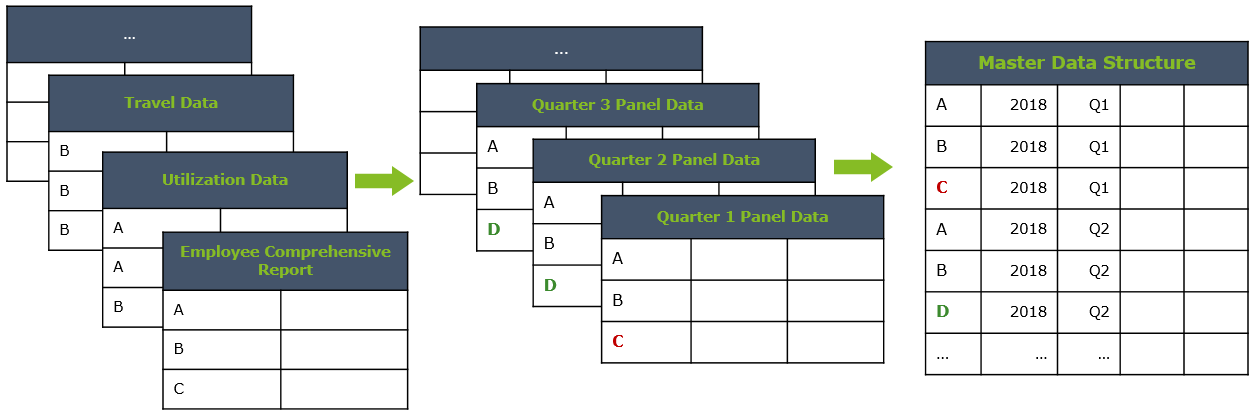


All the data extracts are flat files. General structure of each data source remains relatively similar and there is some common variable like employee ID in all data sets, then the process of joining these together becomes very easy indeed. As more data is added to each of the disparate sources, the code can simply be re-run to quickly re-integrate the data and reproduce new, updated analyses.

**3.2** **Data Structure:**

Employee details and behaviors is captured at each quarter and an employee panel master data is created. This is a crucial step since the accuracy of data analysis insights is highly dependent on the quantity and quality of the data used. Gathering accurate data of high quality and a large enough quantity is necessary to produce relevant results.

Below diagram shows the panel structure of employee master database which is used for the model.



**3.3** **Data Preparation:**

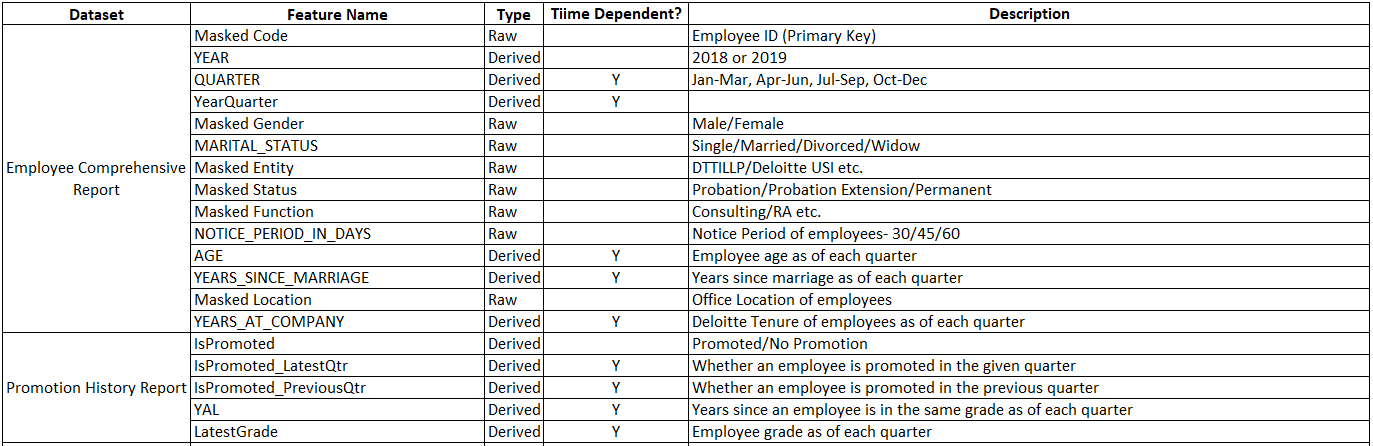
Raw data has been cleansed and transformed to make it usable for the model.

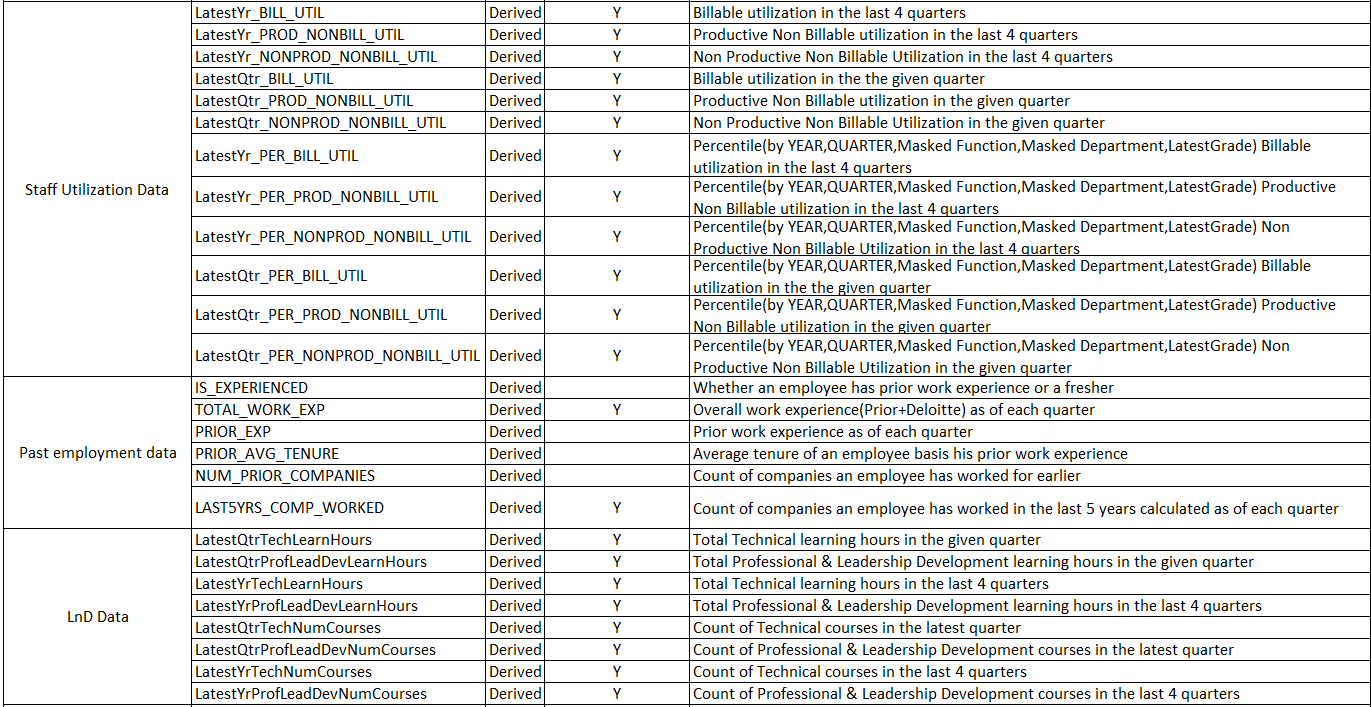
Listed below are the steps taken:

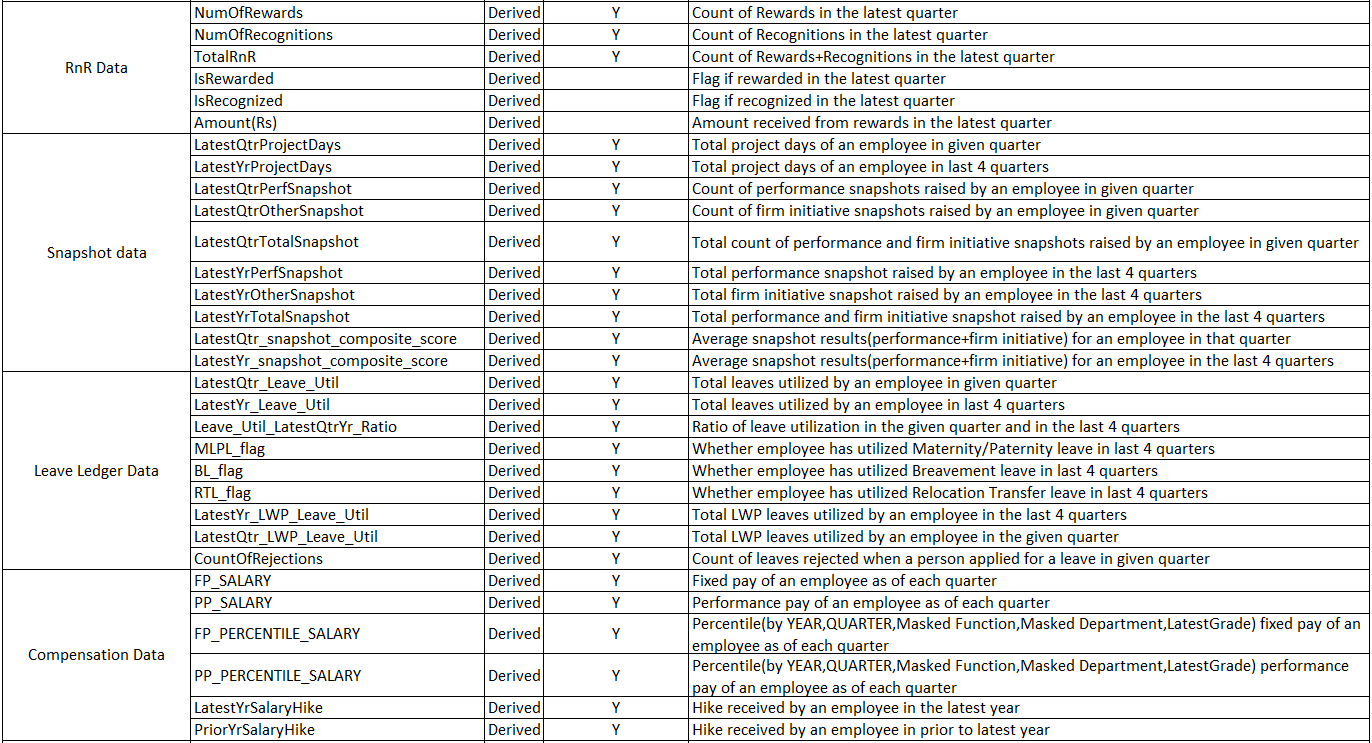
* Using statistics to define normal data and identify outliers.
* Identifying columns that have the same value or no variance and removing them.
* Identifying duplicate rows of data and removing them.
* Marking empty values as missing.
* Imputing missing values using statistics or a learned model.
* Data transforms are used to change the type or distribution of data variables.
* Continuous attribute values are substituted by small interval labels.
* Data reduction mechanism in for some columns as it transforms a large dataset into a set of categorical data.
* New attributes are created from an existing set of attributes.
* Data is normalized so that it falls under a given range.
* Smoothing used to remove noise from the dataset using some algorithms which allows to highlight important features present in the dataset.

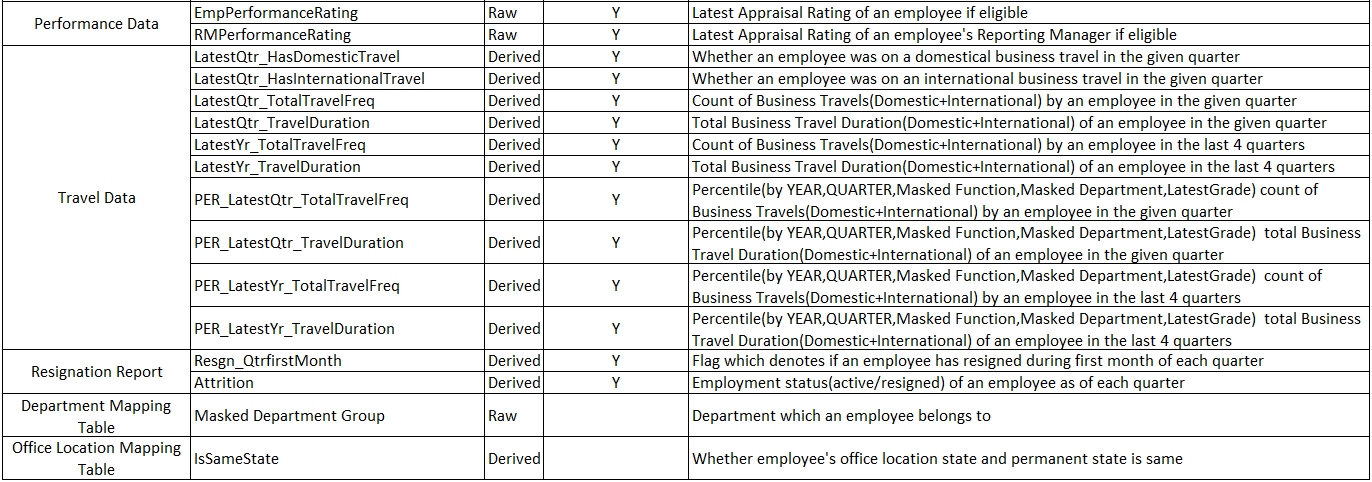
**3.4** **Derived Variables Creation:**

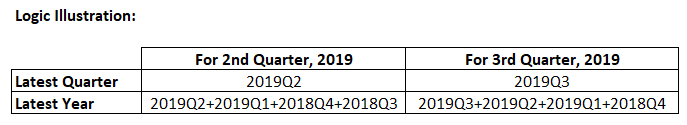
After the data is cleansed, basis the domain understanding, following features has been created.











**4.** **Model Development:**

**4.1** **Defining Model Selection Metrics:**

Sensitivity, Precision and Accuracy are the metrics considered for the evaluation of the model performance.

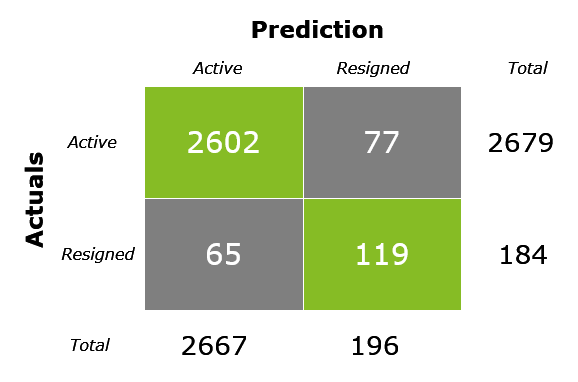
Sensitivity - Ratio between how much were correctly identified as resigned by model to how much were actually resigned.

Precision - Ratio between how much were correctly identified as resigned out of all resignations predicted

Accuracy - Ratio of the number of correct predictions by the number of total predictions

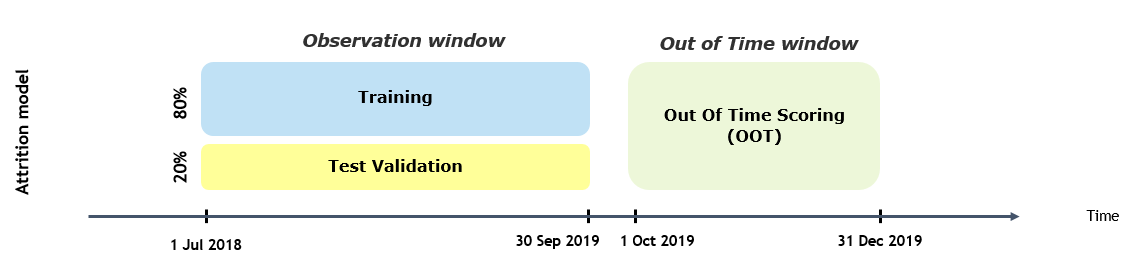
An illustration of model selection matrix:

Total count of employees: 2863



|  |  |  |
| --- | --- | --- |
| **Metric** | **Definition** | **Formula** |
| Sensitivity | Out of Actual resigned employees (184) , model correctly predicted (119) | 119/184 = 65% |
| Precision | Out of resignations predicted by model (196), 119 actually resigned | 119/196 = 61% |
| Accuracy | Out of 2863 employees, model correctly predicted actual status (active/resigned) of 2721 employees | 2721/2863 = 95% |

**4.2** **Data Strategy for Model Development:**



* Observation window is Split into 2 buckets – training and test basis 80:20 rule
* Training dataset – Model is trained basis employee historical data (Jul’18 – Sep’19)
* Test dataset - Dataset on which model is trained model is tested to ensure robustness
* Out of Time dataset – Unknown data on which model is again tested for accuracy
* Period selection of training and testing window is basis the availability of historical data as RnR data for 2017-18 and Performance Data for 2017 is not available

**4.3** **Model Exploration:**

Various algorithms has been tried to find the model which best captures the employee behavior. Few of them are listed below:

* Logistic Regression
* Decision Tree
* Random Forest
* LightGBM
* XGBoost
* Auto ML

Sensitivity, Precision and Accuracy are the metrics considered for the evaluation of the model performance.

**4.4** **Variables Reduction:**

* Variables reduced using Information value, Event rate analysis and VIF. The reduced variables have been further selected basis on variable importance as per XGBoost (from XGBoost package).
* This importance is calculated explicitly for each variables in the dataset, allowing variables to be ranked and compared to each other.
* Generally, importance provides a score that indicates how useful or valuable each feature was in the construction of the model. The more an attribute is used to make key decisions in model development, the higher its relative importance.

**4.5** **Model Refinement:**

Optimizing the models parameters:

Optimization problems can be classified in terms of the nature of the objective function and the nature of the constraints. Special forms of the objective function and the constraints give rise to specialized algorithms that are more efficient. Various techniques has been followed to fine tune the model in order to increase the performance of the best model chosen. They are listed below:

* Grid Search
* Random Search
* Heuristic Methods

**4.6** **Model Validation and Testing:**

* Model Validation and Testing is the procedure of evaluating the wellness of models performance against the real data.
* It is essential that the model validated by considering the aspects and the components before introducing them into the production ecosystem.
* Our Model performance is measured against Test and Out Of Time datasets to check for model stability and model generalization. Where Test data set data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters and Out Of Time data used to provide an unbiased evaluation of a final model fit on the training dataset.

**4.7** **Final Models and Results:**

Using multiple machine learning algorithms, we can get an estimate for how accurate each model may be on unseen data. We need to be able to use these estimates to choose one or two best models from the suite of models that we have created.

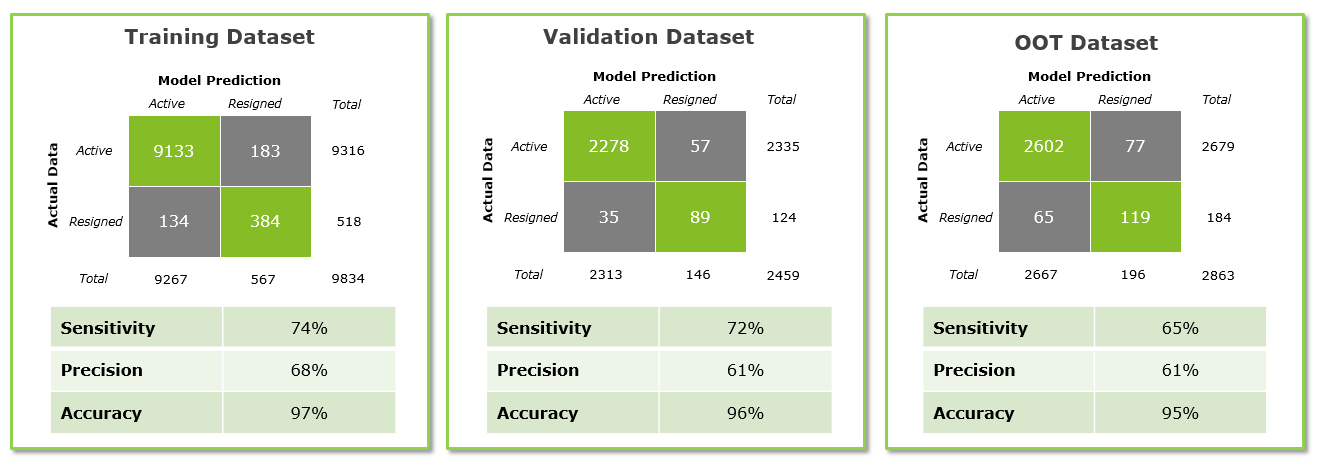
**4.7.1** **Consulting Model:**

Different models has been tried and below is the performance metrics for each of the model.

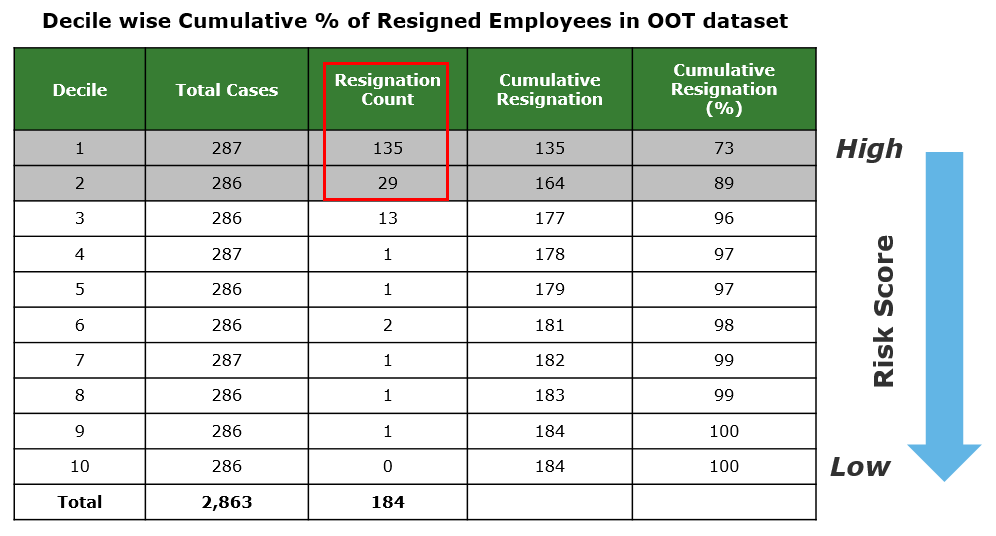
XGBoost performed better in all the three datasets –Train, Test and OOT and has been chosen as the final model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Logistic Regression** | **Random Forest** | **LightGBM** | **Extreme Gradient Boosting** | **Automatic Machine Learning** |
| **Sensitivity** | 59% | 52% | 58% | 65% | 38% |
| **Precision** | 54% | 58% | 64% | 61% | 76% |
| **Accuracy** | 94% | 94% | 95% | 95% | 95% |

Below is the performance values for the XGBoost model across all the datasets.



Below Lift table shows the 89% of the resignee has been captured in the top 2 deciles.



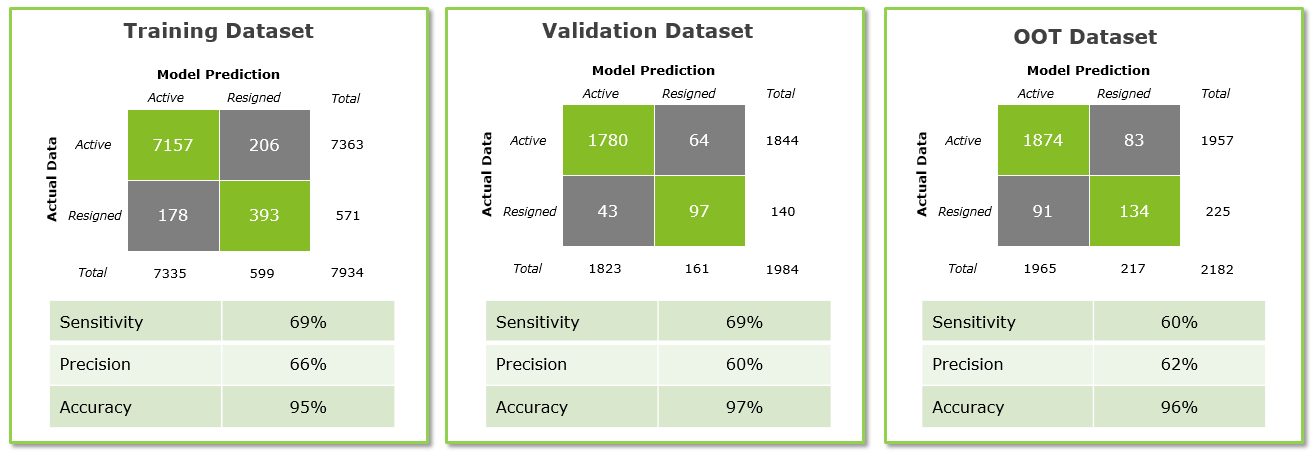
**4.7.2** **Risk Advisory Model:**

Different models has been tried and below is the performance metrics for each of the model.

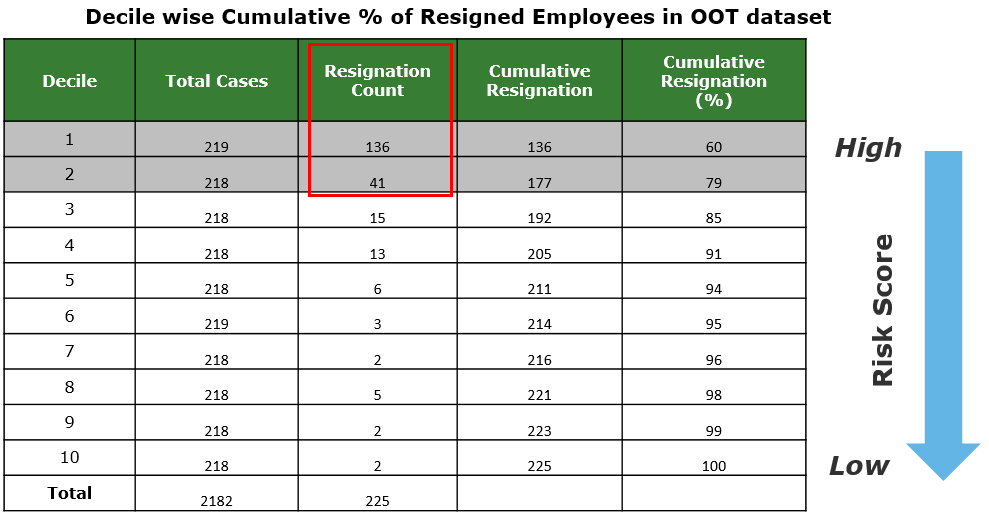
XGBoost performed better in all the three datasets –Train, Test and OOT and has been chosen as the final model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Logistic Regression** | **Random Forest** | **LightGBM** | **Extreme Gradient Boosting** | **SVM** |
| **Sensitivity** | 53% | 48% | 52% | 60% | 50% |
| **Precision** | 70% | 71% | 68% | 62% | 54% |
| **Accuracy** | 94% | 93% | 93% | 92% | 91% |

Below is the performance values for the XGBoost model across all the datasets.



Below Lift table shows the 79% of the resignee has been captured in the top 2 deciles.



**4.8** **Version Control:**

Version control records changes to a file or set of files over time so that you can recall specific versions later.

|  |  |
| --- | --- |
| Step1 | Step2 |
|  | cid:image001.png@01D68DE5.CA83F0C0 |

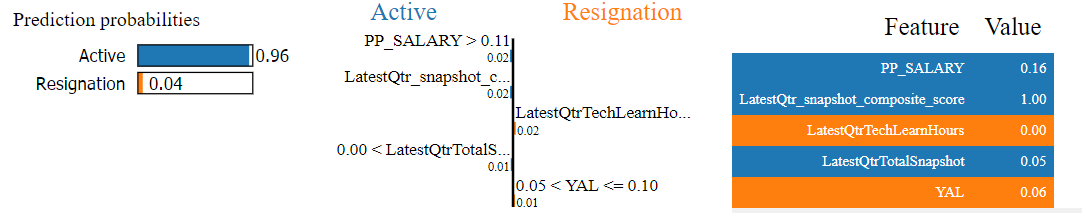
Created a new folder- ‘Code Versions’ to store the different versions of the developed code.

Here we maintained version control manually to edit/update the files and to capture any changes in the code during development since other platforms like GitHub etc. which is a code hosting platform for version control and collaboration works on private repositories.

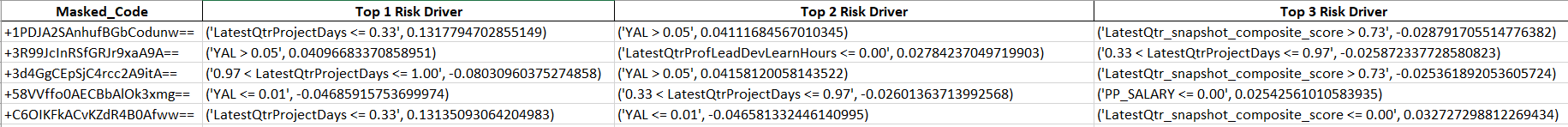
4.9 Lime Explanations:

LIME provides local model interpretability. LIME modifies a single data sample by tweaking the feature values and observes the resulting impact on the output. Primarily using LIME we can predict the probabilities whether an employee will be Active or Resign and also the top N risk drivers which are leading for an employee resignation along with the weightage of the features.

The output of lime looks as below:



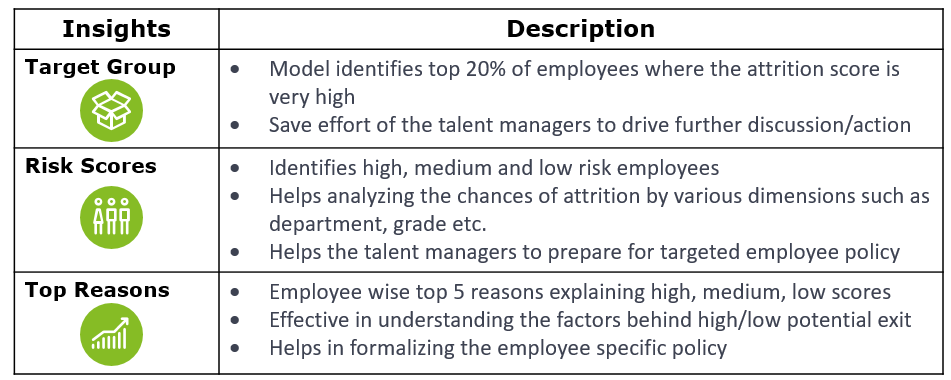
The output seems as below when exported to excel:



**4.10** **Model Output:**

The output of the model helps the talent team to understand the employee groups to target, understand the dynamics of attrition and formulate employee specific policies. The output explains whether an employee is having a high risk or low risk to resign along with the risk scores and its associated top 5 risk drivers.

Model generates below mentioned insights:



Below is the sample output:



5. Model Refresh SOP

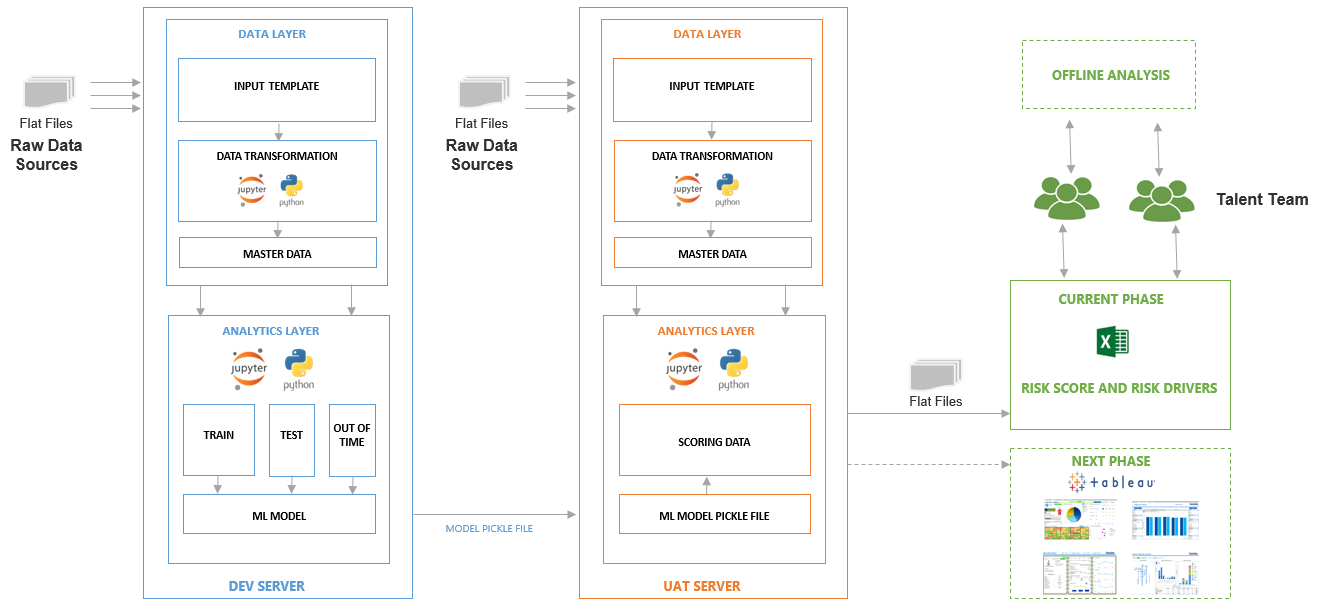
5.1 Architecture:

The current development follow below architecture and the technology used here to develop the solution is Python (3.8.5)/Jupyter Notebook.

The code has been developed on windows server (RAM: 4 GB, Storage: 50 GB)

Server details:

|  |  |
| --- | --- |
| **Server** | **Environment** |
| INMUMCLOUD0422 | DEV |
| INMUMCLOUD0421 | UAT |



Code has been developed on DEV server and productionized on UAT server.

**DEV SERVER:**

DEV server has been used for developmental activities to get the best Model and below are steps followed for the same:

1. DATA LAYER:

* Collected the input data files required for analysis from the Talent team and performed the data quality checks.
* Designed an employee level panel data for each quarter, mapped the datasets and created the master data. Applied the data transformations and treated the outliers, missing values etc.

1. ANALYTICS LAYER:

* Divided the master data based on Function (consulting, RA) an employee belongs to and kept the relevant variables required for the model based on EDA, correlation etc.
* Split the data into train, test and OOT (unseen data). Trained the model and verified the performance on test data. Validated the accuracy on unseen data.
* Built the model using significant variables identified and selected the best model based on the metrics across recall and precision with good accuracy.
* Generated employee Risk Scores and top Risk Drivers using Lime package.
* Stored the best chosen model and the lime explainer as pickle and dill files respectively to be used for scoring in the UAT server.

**UAT SERVER:**

UAT server is used to generate the required output.

1. DATA LAYER:

* Convert the input data as per Input data standard template
* Performs the data quality checks and data transformations
* Generates the employee level data from multiple data extracts

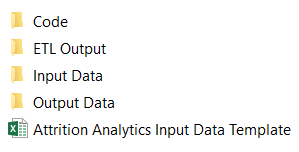
1. ANALYTICS LAYER:

* Loads the saved pickle and dill files for scoring and generating the risk drivers for the newer data
* Generates the required output by Function

As of now the output is in excel format and going forward, the same could be shown in the dashboard (Tableau, Power BI etc.)

5.2 Folder Structure:

Folder structure is maintained with one root folder as ‘Attrition Analytics’ with 4 sub folders in server as per below:



**- Attrition Analytics (root)**

**- Code**

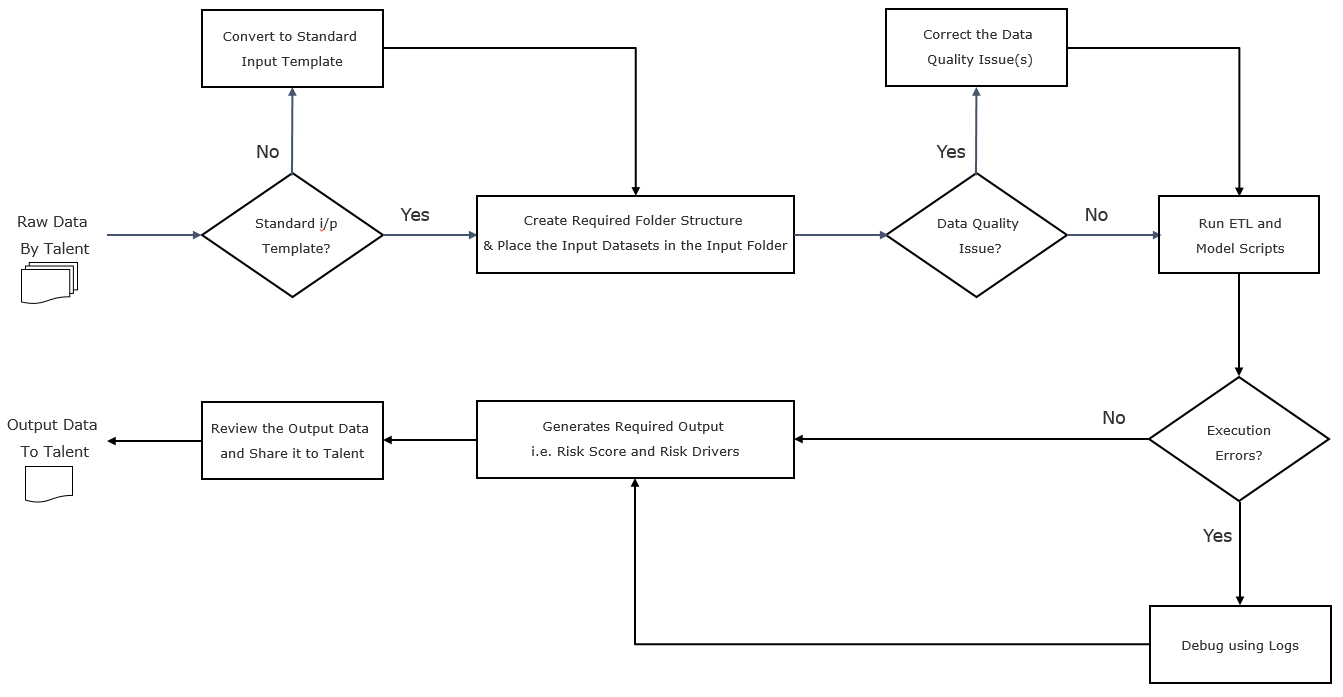
**- ETL Output**

**- Input Data**

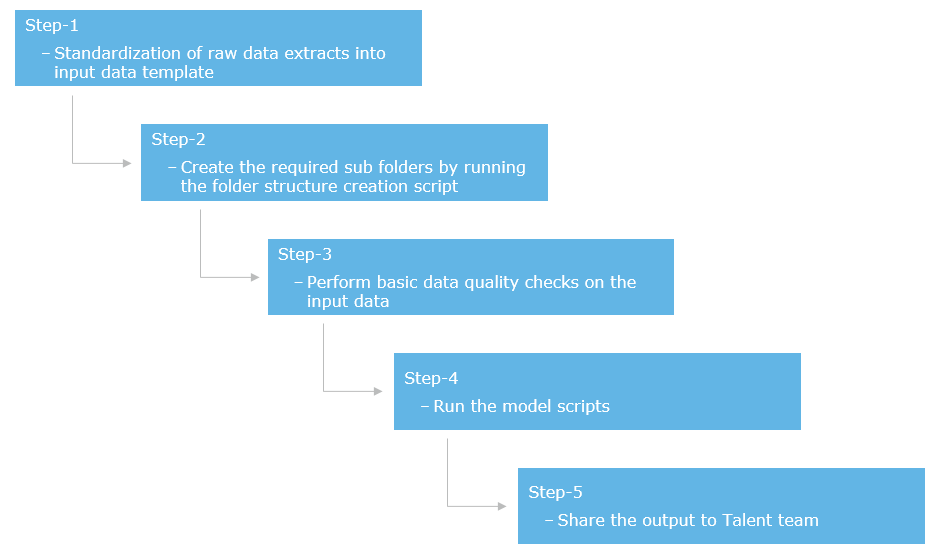
**- Output Data**

All the codes are available in the folder “Code”. Run the “Attrition\_Analytics\_FolderStructure\_Creation.ipynb” script which creates sub folder as “Month Year” in “ETL Output”, “Input Data” and “Output Data” folders and also creates individual input data folders where the input files need to be placed in “Input Data” folder.

**5.3** **Process Flow:**



Below are the standard steps to be followed while refreshing the data in UAT server:



5.3.1 Model Input:

Input data files should be converted as per standard data template (i.e. file name, sheet name, file format, column names etc.) and then files to be placed in the respective folder in order to execute the code successfully.

Attached Input standard template for reference:



|  |  |
| --- | --- |
| Step1 | Step2 |
|  |  |

**Steps to prepare input data:**

**Step-1:** All the sub folders in step1 and step2 as shown in above table will get created as per 5.2

Ex: The step1 folder (“Month Year”) in above table will be given name as on the month and year the script has run and the folders in step2 will get created inside step1 folder.

**Step-2:** Fill data inside respective individual data folder with file name and format as per ‘Attrition Analytics Input Data Template.xlsx’

Ex: Store data in individual folder for all filename with excel format and exact sequence of column names provided in respective tabs with their filename

5.3.2 Model Code Run Procedure and Logs:

Steps to be followed before model run:

**Step-1:** All model related code files are inside ‘Code’ folder. So whenever a user wants to run the model, the files should be run in sequence.

**Step-2:** Run the “Attrition\_Analytics\_ETL.ipynb” file which contains all the data pre-processing steps and these generates ETL output files which will used as input data files to run the model scripts

**Step-3:** Run the Attrition Analytics\_Model\_F002.ipynb and Attrition Analytics\_Model\_F005.ipynb which are the notebook files for Consulting and RA model respectively.

**Step-4:** Logs would be generated and this can be used to debug any errors/view the status of code execution.

**5.3.3 Output Generation:**

The “Output Data” folder will have generated output for both Consulting and RA data after running the model in above step3 in 5.3.2. The output file names are as below:

- Attrition\_Analytics\_F002 for Consulting

- Attrition\_Analytics\_F005 for RA

**6.** **Project Assumptions and Controls:**

**6.1** **Project Assumptions:**

* Considered Resignation Date instead of Date of Leaving
* For Analysis, included only permanent employees (i.e. Partners, Retainers, No Show etc. are not considered)
* Only voluntary resignation is considered ( i.e. Death, Resignation(HR) etc. are not considered)
* For employees having missing data entries for Marital status are considered as ‘Not Available’
* To align with the standard Quarter definition from Deloitte Quarter definition which has a lag of one month of data, we have used the appropriate weightage for the Staff Utilization Reports.
* For employees who are not eligible for performance rating for particular year (ex. new joiners etc.) are been tagged as ‘Not Eligible’
* Capped the number of project days which are exceeding 91 days and 365 days for quarter and year respectively
* For employee’s having negative salary hike and those who are not eligible for the hike, are been capped with 0
* Capped utilization of an employee who are having value greater than 2 to treat the outliers
* For travel data, we considered the input data is having complete information of employee’s travel details as we cannot source the data which were booked personally other than from Deloitte Thomas Cook system

**6.2** **Project Controls:**

**TOLLGATEs** to be checked are added in the code

Data TOLLGATEs

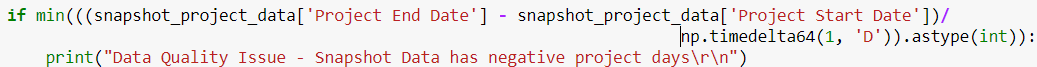
**TG1** - To identify if there are multiple performance groups for employee

Ex:



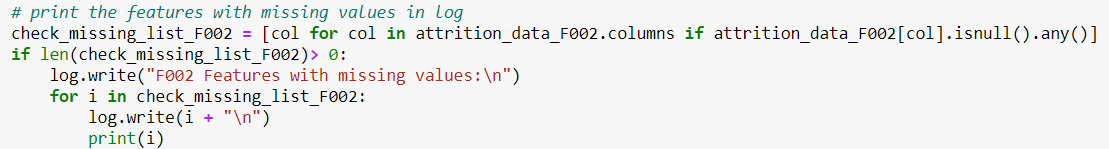
**TG2** - To identify if there are any negative project days which may occur due to data format issue

Ex:



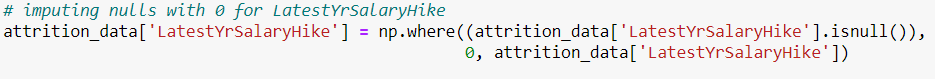
**TG3** - To identify if any variable including in the output file is having missing values

Ex:



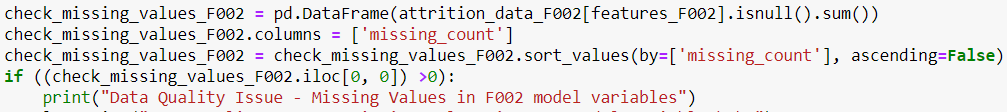
**TG4** - Identified outliers and missing values in the variables are treated with appropriate approach

Ex:



**TG5** - To identify the missing values in the variables which are used in Model Development

Ex:



**TG6** - To check the distribution of the data and decide whether a model need to be re-trained or not

Ex:

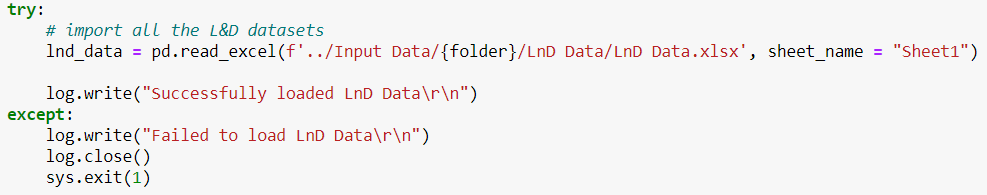


Model run TOLLGATEs

**MTG1** - Added try catch to identify any failures during execution

**MTG2** - Included logs wherever appropriate to capture the status (success/failure) of the execution in a log file

Ex:



About Deloitte

Deloitte provides audit, tax, consulting, and financial advisory services to public and private clients spanning multiple industries. With a globally connected network of member firms in 140 countries, Deloitte brings world-class capabilities and deep local expertise to help clients succeed wherever they operate. Deloitte's 200,000 professionals are committed to becoming the standard of excellence.

Usage Statement

This publication is for internal distribution and use only among personnel of Deloitte Touche Tohmatsu Limited, its member firms, and its and their affiliates. Deloitte Touche Tohmatsu Limited, its member firms, and its and their affiliates shall not be responsible for any loss whatsoever sustained by any person who relies on this publication.

If you will be using this template to prepare document for delivery external to Deloitte, you are required to add the appropriate member firm disclaimer statement.

©2016. For further information, contact Deloitte Touche Tohmatsu Limited.